A Multidimensional Data Warehouse for Community Health Centers

Kislaya Kunjan, MS, MBA[1], Tammy Toscos, PhD[1,2], Ayten Turkcan, PhD[3], Brad N. Doebbeling, MD, MSc[1]

[1]Department of BioHealth Informatics, Indiana University, Indianapolis, IN; [2]Parkview Research Center, Fort Wayne, IN; [3]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA

Abstract

Community health centers (CHCs) play a pivotal role in healthcare delivery to vulnerable populations, but have not yet benefited from a data warehouse that can support improvements in clinical and financial outcomes across the practice. We have developed a multidimensional clinic data warehouse (CDW) by working with 7 CHCs across the state of Indiana and integrating their operational, financial and electronic patient records to support ongoing delivery of care. We describe in detail the rationale for the project, the data architecture employed, the content of the data warehouse, along with a description of the challenges experienced and strategies used in the development of this repository that may help other researchers, managers and leaders in health informatics. The resulting multidimensional data warehouse is highly practical and is designed to provide a foundation for wide-ranging healthcare data analytics over time and across the community health research enterprise.

Introduction

Primary care practices and Community Health Centers (CHCs) are key components of the health care system in the United States (US), providing chronic and preventive care services to low income and medically underserved areas[1]. Incorporating health information technology (IT) systems for healthcare delivery has been recognized as a means to support improvements in patient outcomes, while yielding cost savings for patients, providers and payers[2]. CHCs are atypical in terms of their patient population, which in turn affects their overall clinical operations, including their health IT requirements. With over 96% of the CHCs using EHRs and handling over 90 million patient-visits a year[3], they are an excellent source of clinical and administrative data and an ideal setting for research and development. However, when data resides in multiple disparate silos, the payers and providers cannot aggregate, analyze and assess the data in a cost-effective manner[4]. A data warehouse that can aggregate multiple disparate data sources, including patient records, clinical quality scores, and payer data for CHCs can be used to efficiently improve clinical and financial outcomes across the practice and the region at large[5, 6], as well as to foster informatics and operations research.

In addition to the state-wide operational health information exchanges (HIEs) in the US,[7] there exist several large data repositories/warehouses of significance, some of which include: Healthcare Cost and Utilization Project (HCUP)[8]; National Patient-Centered Clinical Research Network (PCORnet)[9]; Center for Medicare and Medicaid Claims Data[10]; US Food and Drug Administration's Sentinel Initiative[11]; CancerLinQ[12]; and eMERGE (Electronic Medical Records and Genomics)[13]. Enterprise data warehouses to manage heterogeneous biomedical data have been developed for medical centers, such as Veteran Health Administration's Corporate Data Warehouse[14], Intermountain Healthcare's CIS[15], Massachusetts General's COSTAR[16] and Mayo Clinic's Enterprise Data Trust[17].

A CHC data warehouse that combines the best features of the HIEs and consolidated repositories like HCUP, along with health data analytics, can be particularly desirable for CHCs. In this paper we share our journey to create a multidimensional clinic data warehouse for seven clinics in the state of Indiana with whom we are collaborating to improve access to care for underserved patients. The primary research question guiding this work is how to design a CDW to support improvements in care delivery at community health centers.

Background

Indiana is home to the Indiana Health Information Exchange (IHIE)[18], which is the largest HIE in the US. However, Regional Health Information Organizations (RHIOs) that manage such local HIEs have limited utility for CHCs because the participating facilities that share data into HIEs are typically hospitals and integrated health centers, in contrast to community clinics that are free standing and resource constrained. There are additional challenges with

HIEs such as: high start-up (upwards of $12 million) and operating costs[19]; challenges in their business model,[20] need for a dedicated infrastructure to managing streaming data from the healthcare facilities, while addressing issues of privacy and security,[21] creating and maintaining a master patient index or record locator service and finding technologically capable exchange partners[22]; and dealing with issues of distrust and data control issues among competitors[23, 24]. Furthermore these HIEs are primarily focused on patient-identifying clinical data derived from EHR and ancillary systems with a focus on allowing data to follow patients across delivery settings. They are not designed to be data stores or a data warehouse which capture operational and financial data and/or provide data analytics to members in the exchange network.[25] In our research, we discovered that the CHCs in Indiana were not part of the IHIE member network, not much unlike the 280 other statewide operational HIEs in the nation who lack CHC involvement[26]. Most CHCs we interviewed expressed keen interest in being able to effectively draw insights from their data, even as they encountered challenges in doing so. We chose to develop a unique community clinic data warehouse that would capture not only the medical record data but also include operational and financial data of the clinics in a very cost-efficient manner. To this multidimensional data warehouse, we intend to add an analytics platform, with a goal to support better patient care which can be defined in terms of improved access to care, shared decision making, and greater patient satisfaction[27, 28]. In addition, the intent is to enhance the operational and financial health of the community health care systems in the state of Indiana. We also believe there is great opportunity to foster community informatics research and development using this data to improve care delivery, patient engagement and efficiency.

The proposed work is an offshoot of an ongoing 3-year research project funded by the Patient Centered Outcomes Research Institute (PCORI). The larger project aims to improve access to care for CHCs by: identifying common barriers and successful patient-centered strategies; applying Delphi expert panel methods along with workflow and simulation modeling to identify feasible and valid patient-centered, strategies; and finally determining if implementing patient-centered systems can improve access to care, relative to existing approaches. The entire study, including the data collection from the clinics, has been IRB approved by the Indiana University Office of Research Administration.

Methods

We partnered with 7 CHCs geographically spread across the state of Indiana. The clinics were selected such that they represented urban, semi-urban and rural settings and provided care to diverse patient population in terms of race, ethnicity, insurance status, and income levels.

Design of the Community Clinic Data Warehouse

A common and widely accepted definition of a data warehouse is that it is a subject-oriented, integrated, non-volatile, and time-variant collection of data to support decision making[29]. Data warehousing methodologies share a common set of tasks, including business requirements analysis, data design, architectural design, implementation and deployment[30]. We present a data warehouse architecture that is simple, low cost, and one that can readily benefit resource constrained CHCs.

Our multidimensional **Clinic Data Warehouse (CDW)** captures clinical and operational data from community health centers in a uniform format to facilitate comparisons and analyses over time. There are four categories of data acquired from the clinics: 1) Clinic Operational Data; 2) Clinic Financial Data; 3) Clinic Quality Measures Data; and 4) Patient Level Data. These four data categories are captured and integrated into a common CDW. Downstream data analytics is applied to the data to create customized analytic reports of use to the clinic. Separately, patient analytics provide insights into optimizing care for the CHC's patient population.

The information flow diagram for the CDW and associated analytics is illustrated in Figure 1.

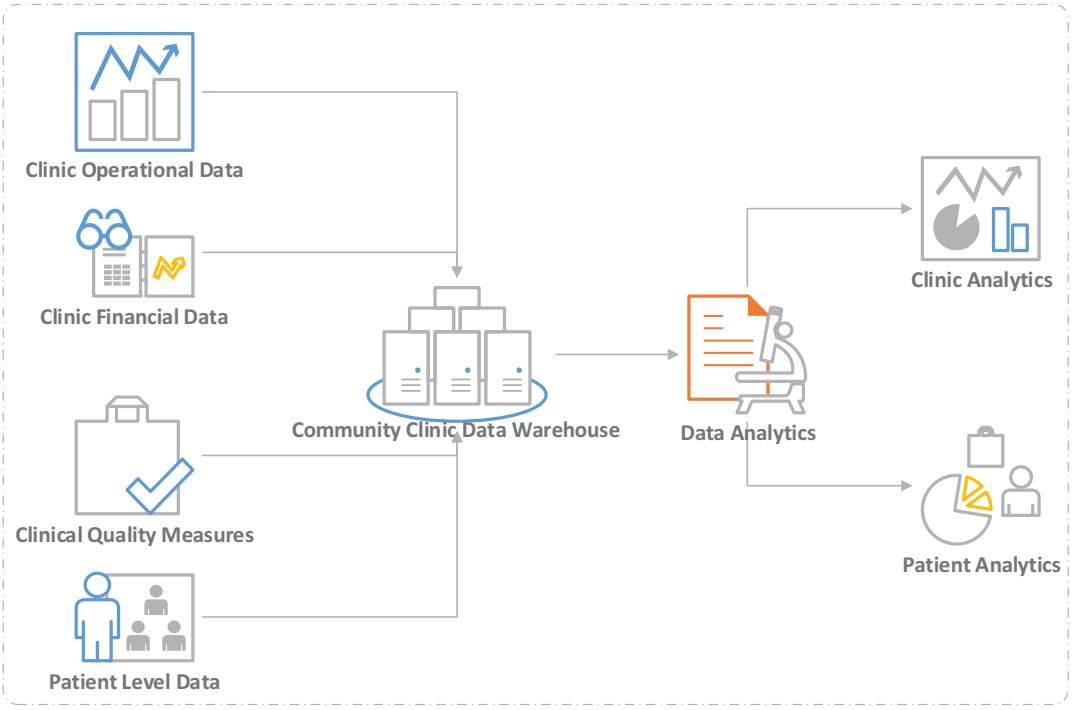

Figure 1: Information Flow Diagram for the Clinic Data Warehouse

The conceptualization of the CDW was based on the requirements gathered from all the stakeholders involved by conducting structured interviews with key informants that included clinic management, staff, providers, leadership, patients, as well as IT and QA/QI staff. While the initial choice of the data elements was driven by a narrow purpose to model and simulate clinic's scheduling system, it soon expanded in scope to serve as a multidimensional repository sourcing information from various domains. The objective is to understand the implications of the clinic operational characteristics on the clinic performance and even patient outcomes over time. For example, the database will allow the study of the impact of scheduling method (such as open access) on appointment compliance (no-shows, cancellations, reschedules), staff utilization, and various patient outcome measures. The system architecture was defined in terms of: data sources, normalized common data model, interoperability requirements, data standards, data acquisition technologies, modes of acquisition, storage in a database server, privacy and security standards set up, data validity checks, data back-up performance and potential to build tools for data management and visualization. Roles and responsibilities were set for each of the above tasks, along with defining the business model.

Obtaining Clinic and Patient Level EHR Data

The clinic data category includes: Operational Data (such as, clinic type, certification status, locations, size, patient mix, payer mix, service mix, staff utilization, scheduling system, enrollment system, quality measures and other dimensions of care); and Financial Data (revenue, expense, margins, cash flow, payroll). The clinic operational/financial data was obtained from the audited financial statements of the clinics, IRS Form 990 (annual returns for certain federally tax-exempt organizations) and Uniform Data Systems (UDS) filings. This was supplemented with a structured questionnaire completed through an interview with the clinical manager, followed with workflow observations and opportunistic interviews as appropriate, in the clinic.

The patient level data serves as a backbone of the data warehouse. The data elements selected are those that are digitally captured as part of routine clinical practice and reported to federal agencies. Collecting this data required the design of a common data model, given that we were collecting information from different healthcare facilities using different Patient Management and EHR systems. The common data model helped promote efficient and streamlined collection of the data generated in each of the healthcare centers by organizing data into a standard structure. We developed a common data model that incorporated elements and practices of those used by other large organizations, including: HL7's C-CDA, PCORnet and FDA's Sentinel Initiative.

While we explored multiple ways to acquire patient data from the EHR, the one we selected was dependent on the EHR reporting capabilities, IT resources in the clinic, and the data architecture of the CDW. We realized that it was most efficient to obtain the data dictionary/metadata (tables, column names, definition/meaning, data type, predefined values sets/descriptive text) of the EHR where possible and identify the various data elements of our interest. The data elements were selected from well over 3000 tables present in the data dictionaries of the respective CHCs. We then generated SQL queries for each of the individual EHRs and gave them to the clinic staff to execute them on their database server. Once the data was extracted and the de-identification was verified in accordance with HIPAA safe-harbor provisions[31], the clinic shared the data with us through secure email. The raw data was prepped programmatically before populating into our Microsoft SQL Database Server based on the common data model of the CDW and made ready for ad-hoc querying/reporting. Figure 2 shows this process flow of data collection from the clinic EHRs.

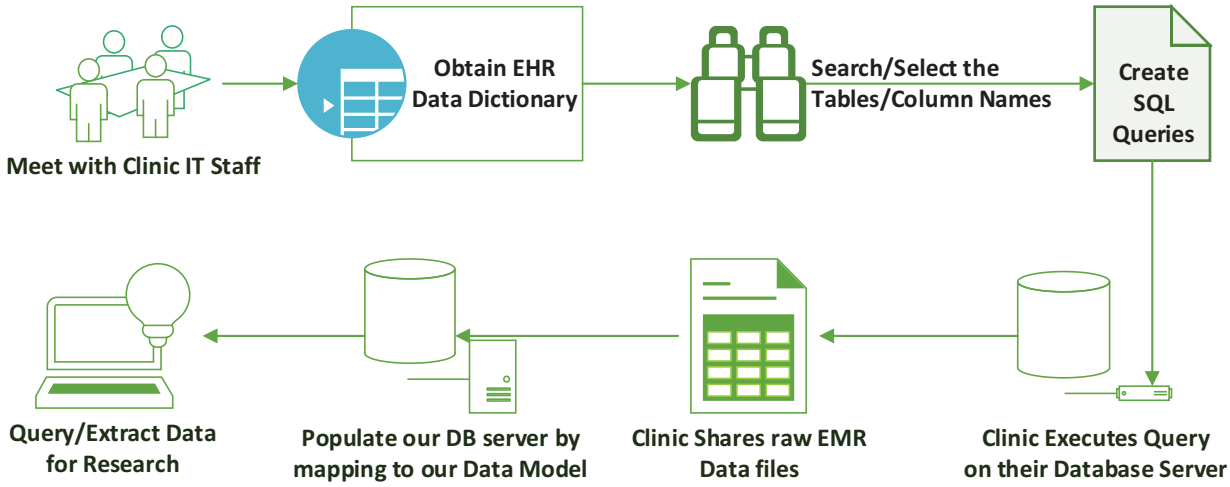

Figure 2: Process flow of data collection from the Clinic EHR Systems

The core data element categories included: patient demographics, appointment data, date/time stamps of key events for patient and data flow, insurance information, provider information, medical problem list, encounter diagnoses, immunizations, medications, laboratory data, care plan, procedures, and healthcare referral information**.**

Results

We now present the results in terms of the common data model developed and a selection of graphs to illustrate the content of the CDW.

Common Data Model

A common data model for the CDW has been built which is guiding the collection of specific data elements. The majority of the data collection exercise, comprising up to five years of EHR and UDS data from all participating clinics, has been completed. For all subsequent data requests (which will be done on a quarterly basis to begin with), the clinic will only need to re-run the queries to append to the current database housed at our end. We mapped the data to the same consistent format (e.g., with the same variable name, precision, and other metadata). By undertaking this step, we created a platform that enables much more rapid analytic capabilities. The data model is expected to grow and evolve over time by including additional data elements that are analytically important.

Note that the data was already de-identified before populating into our database. A pseudo patient ID was used to map the data residing in different tables. In the first round of data collection, many of the raw data tables had to be prepared before loading into our database. Populating free text information was particularly challenging given the existence of delimiters like commas, tabs and new lines. But this was resolved programmatically and by specifying the input file format. Future data loading will be much more streamlined. The main tasks in developing the common data model were the selection of appropriate entities/tables and attributes as well as data normalizations based on the three different commercial EHR systems we encountered.

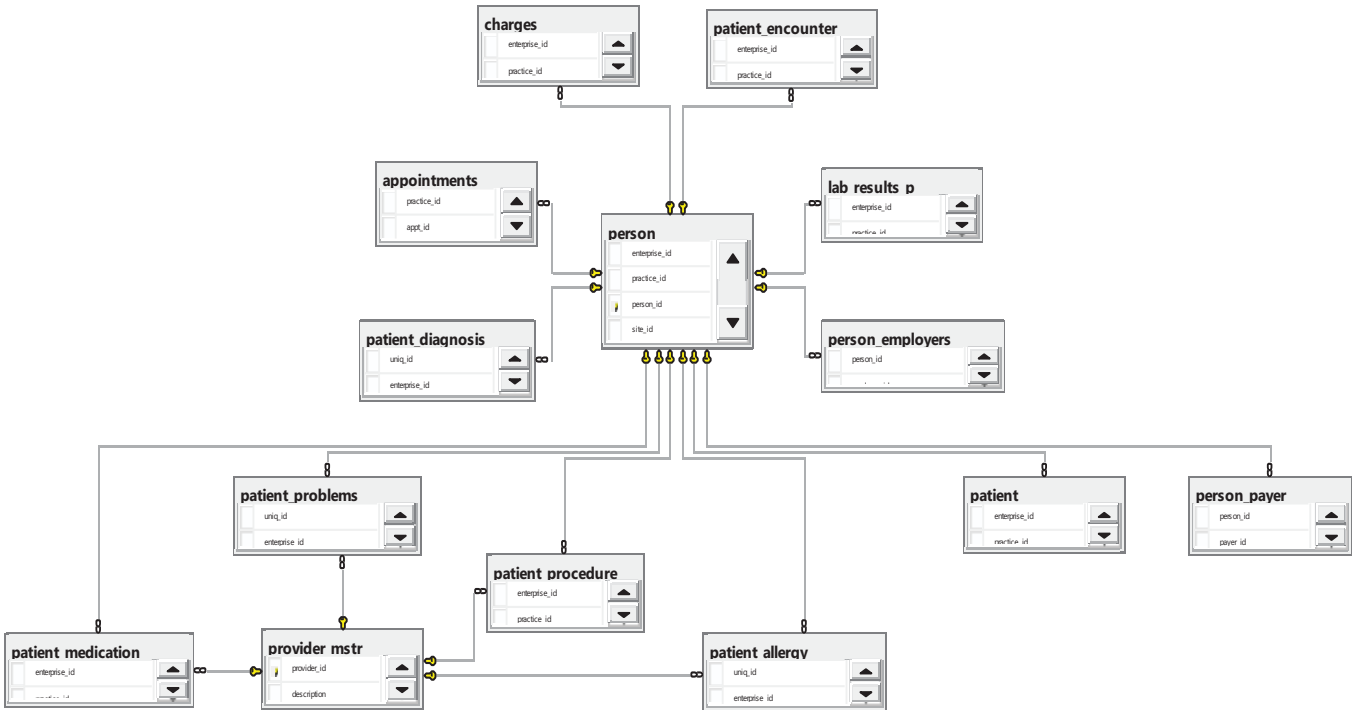

Figure 3: Snapshot of the truncated E-R/Database Diagram of the CDW

The clinic data warehouse currently stores over **3 million** patient-visit records from 2010-14 from the different CHCs and includes detailed information on appointments, prescriptions, lab tests, screening/preventive measures, immunization, and quality measures under Meaningful Use (MU) Stage 2 and UDS. This is a significant volume of rich longitudinal data that will readily lend itself to large-scale data analytics and operational and efficiency studies.

Select Measures on the Data Collected

We present here a partial selection of the key data measures graphically to illustrate the data content.

Figure 4 describes the racial distribution and insurance status of the patient population comprising each of the 7 Indiana CHCs. The distribution clearly shows that the clinics are not homogenous in their patient population and associated needs.

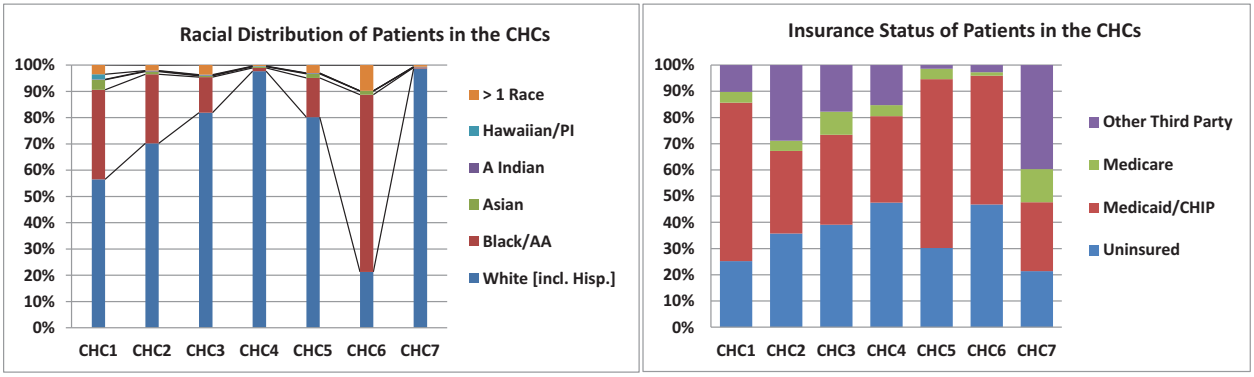

Figure 4: Summary statistics on race and insurance Status

Referring to the plots, it may be noted that the 'White' race includes Hispanic/Latinos, since the latter is not defined as a race, but as an ethnicity. In these CHCs, the percentage for Hispanic/Latino ethnicity ranges from a low 1.2% to high 38%. For insurance, a vast majority of them are either uninsured or have Medicaid, with income levels below

200% of the Federal Poverty Line. Yet, there is a sizeable percentage of patients with private insurance in couple of the clinics.

Among our seven partnering CHCs, four use traditional scheduling, while three use modified open access (a combination of same day open slots and traditional/advanced booking). High patient appointment no-show rates and the associated variability in workload due to prevalent practice of overbooking remains a significant challenge in all CHCs. For example, after analysis of the appointment data in the CDW for one of the CHCs, we found that their no-show rate for 2014 was **21%** and the appointment cancellation rate was **13%**. Providing clinics with easy access to this data and a means to benchmark their performance against other similar organizations can be powerful and create opportunities for design and testing of new technology solutions and positive change.

For another large CHC, we analyzed the visit statistics, in terms of clinic services utilized, and the top 6 disease diagnoses and the associated annual visit frequencies (see Figure **5**). It is important to understand the most common medical conditions and service patterns for these predominant conditions in order to focus efforts on study of these problems, and identify potential technology tools that might help improve these challenges. Focus on highly prevalent problems is likely to have the greatest impact on the overall clinic efficiency and other operational and financial measures.

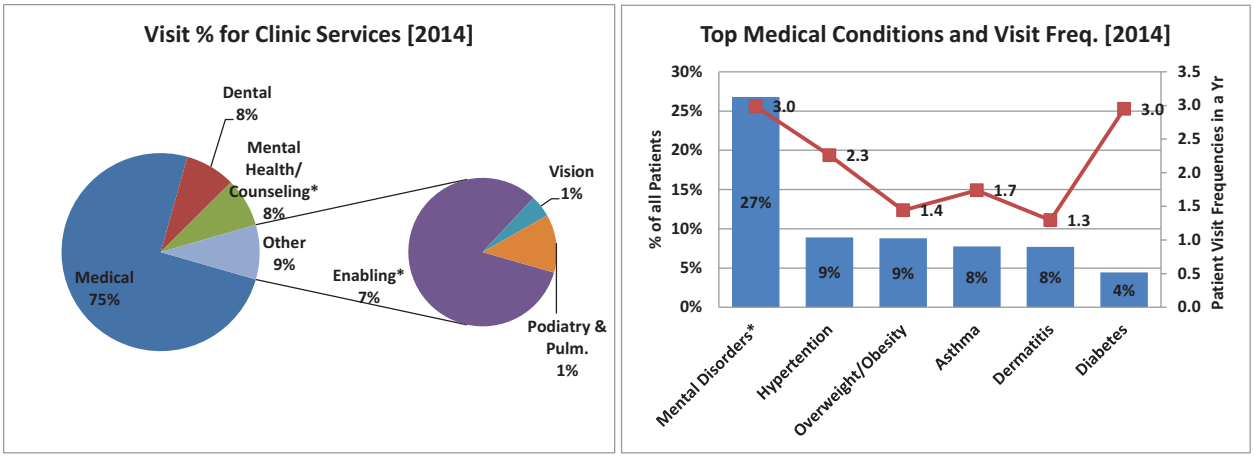

Figure 5: Statistics on services utilized[1] and medical conditions with associated visits[2] at a large CHC in Indiana

Discussion

We have established a novel information exchange/clinic data warehouse that will enable data-driven research, technology development and testing, and informed decision making over the years. We have presented only a small selection of the data in the repository for illustration and plan a number of studies to better investigate its use and applications. We believe this approach can provide an efficient and effective means for CHCs to leverage data to improve patient access and the quality of care delivered to underserved patients.

Academic Research and Benefits to CHCs

There are innumerable research opportunities that can come from analytic study of the CDW. These may be broadly classified as being, **descriptive, predictive, and prescriptive analytics**. A few examples under, 'descriptive analytics' are: development of enriched careflow/process models based on EHR time-stamps; ranking clinics based on adherence to evidence based best practices; and peer comparisons at the state and national level. Under 'predictive analytics' the plan is to build machine-learning models of patient no-show behavior and treatment compliance; to perform agent-based modeling and simulation of clinic operations to improve efficiency; and to perform longitudinal network analysis and risk stratification of diseases. Note that although the patient data in the CDW is deidentified, a pseudo patient ID will still allow us to conduct patient-centered longitudinal studies across

[1] Enabling services in the pie chart include: case management, outreach, patient/community education, transportation, eligibility assistance, and interpretation services

[2] Mental Disorders in the graph is cumulative of 4 mental disorder types: Depression, Anxiety and PTSD, Attention Deficit Disorder, and Other Mental Disorders.

care settings. Assessing the impact of the affordable care act (ACA) on access to care will be of particular and timely relevance.

Under 'prescriptive analytics' we intend to incorporate findings from predictive analytics into standalone software applications or embed within patient management systems to provide personalized decision support based on patient profile and clinic characteristics. The data can be further enriched by including external variables, such as data from community/geo-spatial information systems. We also are particularly excited about the opportunity for research and development of new technologies for patient engagement and management in the community, as well as integration of disparate care providers and care sites (inpatient, outpatient, community, "minute" clinics, prisons, other) as part of a care management team. The community clinics stand to benefit from such analytics since they lack the infrastructure that is usually available to academic medical centers.

Ongoing Work

We are refining the design of a data-driven clinic dashboard that will be linked to the CDW. **Figure 6** shows the mock-up of this dashboard under development, which initially will focus on measures of access to care and efficiency. It has a selection of key metrics across different categories (operational, quality, and financial) and has been derived from key informant interviews. It will allow clinic management to visually assess the impact of their operational changes (such as scheduling, staffing level, and policy changes) on clinical and non-clinical outcomes on a periodic basis.

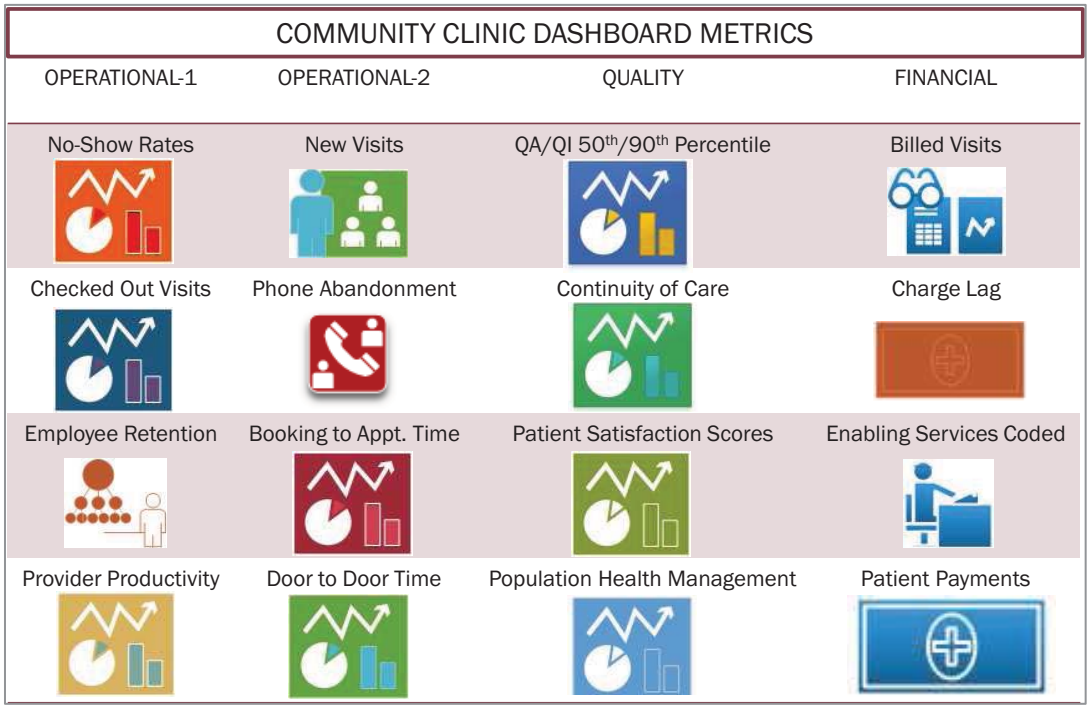

Figure 6: Mock-up of a community clinic dashboard with associated metrics of relevance

We are also working on a secure web portal to serve as an interface for clinic data submission and running queries by researchers. The CDW portal will allow users to explore, interact with, and export data through charts, maps, reports, analyzer and locator tools, data downloads and data services, and widgets. Finally, we are developing strategies for interacting with technology developers and entrepreneurs, in order to test the feasibility of new tools and prototypes.

Challenges and Strategies

There were various challenges encountered in acquiring the data from the clinics. For most clinics, this was the first time they received such a request to extract EHR data. Thus far, the clinics primarily ran reports to fulfill government requirements, in tandem with manual charting. Even though the clinics lacked adequate time or

resources for learning and drawing insights from their data, they all readily recognized its importance to support their evidence-based quality improvement efforts, and be in compliance with the upcoming MU Stage 3 requirements and intent to seek PCMH certifications.

In terms of strategies, we sought to engage meaningfully with all the clinics and presented our value proposition to the clinic leadership. We interviewed the clinic staff and patients to understand their challenges and learn from their innovations, coupled with onsite workflow observations of key processes. We were able to obtain the EHR data dictionary from 2 different EHR systems (covering 5 out of the 7 clinics). These EHRs were based on client-server architecture. The data dictionary allowed us to identify all the data elements of interest to us and create the SQL queries to be executed on their database servers. This made it less onerous on the clinics as they were able to run the provided queries and share the extracted data with us in a secure mode. The remaining 2 CHCs were using a cloud-based EHR with no direct access to the database server. We had to work rather closely with them to pull the data from their web-based reporting tool.

Study Limitations

There were a few limitations to our study. First being, that since the data dictionaries were from three different commercial EHRs vendors, the common data model may not reflect the universe of EHRs in the ambulatory care space. Furthermore, the data warehouse architecture has been partially influenced by the unique characteristics of CHCs in the state of Indiana and therefore may not reflect the particulars of other CHCs in the US. The model may be even less representative of other primary care practices who are not dedicated to meeting needs of the underserved. Finally, the design of the architecture and resource constraints imposes certain practical limitations on the frequency of data collected from the clinics, which in turn would define the type of analytics and decision support that can be performed.

Conclusions

We have described a multidimensional clinic data warehouse for CHCs spread across Indiana and one that integrates operational, financial, quality and patient data. We chose a centralized data architecture that is relatively easy to implement and sustain, yet is rich in scope and content, and relevant to the participating clinics. The goal is to enable free standing CHCs to participate in this data warehouse, in the same way that large hospitals participate in HIEs and benefit from large scale healthcare data analytics to improve overall clinic operational performance and achieve better scores on the ever evolving clinical quality measures. Developing strategies for interacting with technology developers and entrepreneurs, in order to test the feasibility of new tools and prototypes should speed innovation and dissemination of effective IT. Ultimately, we believe that providing the CDW will help CHCs better meet the needs of their patients.

Acknowledgement

The author would like to thank and acknowledge PCORI, which has funded this 3 year, $2M grant award [IH-12-11-5488; 'Improving Access to Care and Efficiency of Healthcare Systems for Underserved Patients', PI: Doebbeling]. The author would also like to acknowledge the staff at the participating clinics for their support and the IU School of Informatics and Computing for providing the necessary technology infrastructure for the implementation of this clinic data warehouse.

References

[1] Shi, L., & Stevens, G. D. The role of community health centers in delivering primary care to the underserved: experiences of the uninsured and Medicaid insured. The Journal of ambulatory care management 2007;30(2):159-170.

[2] Buntin, M. B., Burke, M. F., Hoaglin, M. C., & Blumenthal, D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health affairs 2011;30(3):464-471.

[3] HRSA Health Center Data: http://bphc.hrsa.gov/healthcenterdatastatistics/index.html; Accessed Feb 23, 2015

[4] Sukumar, S. R., & Ferrell, R. K. 'Big Data'collaboration: Exploring, recording and sharing enterprise knowledge. Information Services and Use 2013;33(3), 257-270.

[5] McGinnis, M., Olsen, L., & Goodby, A. W. Clinical Data as the Basic Staple of Health Learning:: Creating and Protecting a Public Good: Workshop Summary. National Academies Press. Chicago 2010

[6] Ledbetter, Craig S., and Matthew W. Morgan. "Toward best practice: leveraging the electronic patient record as a clinical data warehouse." Journal of Healthcare Information Management 2001;15.2:119-132.

[7] Yoder, D.. Health Information Exchange. Information & Communication Technologies in Healthcare, 25. 2011

[8] NIS, HCUP Nationwide Inpatient Sample. "Healthcare cost and utilization project (HCUP)." 2007.

[9] PCORnet: http://www.pcornet.org/; Accessed Feb 30, 2015

[10] CMS Data Navigator: http://dnav.cms.gov/; Accessed Feb 30, 2015

[11] Robb, M. A., Racoosin, J. A., Sherman, R. E., Gross, T. P., Ball, R., Reichman, M. E., ... & Woodcock, J. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. Pharmacoepidemiology and drug safety 2012;21(S1):9-11.

[12] Sledge Jr, G. W., R. S. Miller, and R. Hauser. "CancerLinQ and the future of cancer care." American Society of Clinical Oncology educational book/ASCO. American Society of Clinical Oncology. Meeting. 2012.

[13] Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., ... & eMERGE Network, T. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. Genetics in Medicine 2013;15(10):761-771.

[14] Fihn, S. D., Francis, J., Clancy, C., Nielson, C., Nelson, K., Rumsfeld, J., ... & Graham, G. L.. Insights From Advanced Analytics At The Veterans Health Administration. Health Affairs, 2014;33(7), 1203-1211.

[15] Clayton PD, Narus SP, Huff SM, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med 2003;42:1–7

[16] Murphy SN, Morgan MM, Barnett GO, et al. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp 1999:892–6

[17] Chute, Christopher G., et al. "The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data." Journal of the American Medical Informatics Association 2010; 17.2:131-135.

[18] Indiana Health Information Exchange: http://www.ihie.org/; Accessed Feb 30, 2015

[19] Grossman, J. M., Kushner, K. L., November, E. A., & LTHPOLICY, P. C. Creating sustainable local health information exchanges: can barriers to stakeholder participation be overcome? Washington, DC: Center for Studying Health System Change 2008.

[20] Mcllwain, K., & Lassetter, J. S. Building sustainable HIEs. Health Management Technology, 2009;30(2), 8-11.

[21] Centers For Disease Control and Prevention. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. MMWR: Morbidity and Mortality Weekly Report, 2003;52(Suppl. 1), 1-17.

[22] Vest, J. R., & Gamm, L. D. Health information exchange: persistent challenges and new strategies. Journal of the American Medical Informatics Association, 2010;17(3), 288-294.

[23] Terry, K.. The rocky road to RHIOs. Medical economics, 2006;83(4), TCP8-TCP10.

[24] McGee, M. K. Urgent care. Information week, 2007;28:40-50.

[25] Adler-Milstein, Julia, David W. Bates, and Ashish K. Jha. "Operational health information exchanges show substantial growth, but long-term funding remains a concern." Health Affairs 2013;10-1377.

[26] Callan, Kathy, et al. "Tracking HIE's Ever Evolving Operational Models: Emerging health information exchange market still sorting out its business and governance models." Journal of AHIMA 85.10; 2014:36-40.

[27] Gerteis, M. Through the patient's eyes: understanding and promoting patient-centered care (Vol. 531). Jossey-Bass Inc Pub. 1993

[28] Nuss, T. D. Patient-centered care. Achieving STEEEP health care: Baylor health care system's quality improvement journey, 2013;137.

[29] Inmon, William H. Building the data warehouse. John wiley & sons, 2005.

[30] Sahama, Tony R., and Peter R. Croll. "A data warehouse architecture for clinical data warehousing." Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68. Australian Computer Society, Inc., 2007.

[31] US Department of Health and Human Services. "Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." 2013